2005 한국경영과학회/대한산업공학회 춘계공동학술대회
2005년 5월 13일~14일, 충북대학교

오차패턴 모델링을 이용한 지도학습 모형에서의 성능 향상

Improving the Performance of Supervised Learning Models using Error Pattern Modeling

허 준*, 김종우**
*SPSS Korea (주)Data Solution
**한양대학교 경영대학 경영학부

Abstract

본 논문은 이분형 목적변수를 가지는 데이터에서, 의사결정나무나 신경망과 같은 지도 학습(Supervised Learning)의 훈련을 통한 각종 예측 및 분류 정확도를 향상시키기 위해서 오차 패턴을 이용한 새로운 Hybrid 데이터 마이닝 기법을 제안한다. 오차 패턴을 이용한 Hybrid 기법이란 데이터 마이닝의 서로 다른 기법을 각 데이터에 적용한 다음 기법 간의 불일치되는 부분만을 다시 패턴화 하여, 이를 최종 모형에 적용하여, 기존에 1개의 방법만을 사용하였을 경우보다, 더욱 좋은 정확도를 가질 수 있도록 하는 방법이다. 본 기법의 검증을 위하여, 10개의 실제 검증용 자료를 사용하였으며, 분석 결과 신경망과 의사결정나무 분석과 같은 기존의 방법보다 전체적으로 예측력이 향상됨을 보였다.

1.서론

데이터 마이닝(Data Mining)의 가장 알려진 모델 중 하나는 지도학습(Supervised Learning)기법이다. 대표적으로 MLP(Multi-Layer Perceptron), RBF(Radial Basis Function) 알고리즘을 이용한 신경망 분석(Neural Networks)이나 C4.5, C5.0, CART, CHAID 알고리즘을 이용한 의사결정나무 분석(Decision Tree Induction)이 대표적인 지도학습의 분석 알고리즘이다. 지도학습 기법에는 이외에도 다른 분석 기법들이 존재하지만 위에서 열거한 분석 알고리즘들이 보편적으로 가장 많이 사용되거나 잘 알려진 기법이다. 이런 지도학습의 활용에 있어서 중요한 질문 중에 하나는 "예측율 또는 정확도가 가장 높은 기법이 무엇인가?" 라는 것이다. 데이터의 유형에 따라서 또는 각종 상황에 따라서 가장 좋은 기법이 어떤 것인지에 대한 몇몇 연구된 사례가 있기는 하지만[18], 절대적으로 최선의 방법이 일반적으로 존재하지는 않는다. 즉, 신경망 분석 기법이든 의사결정나무 분석 기법이든 모두 자체적으로 장점과 단점을 가지고 있다. 이렇게 분석기법이 장점과 단점을 가지고 있다는 것은 다시 말하면 특정한 한 개의 기법만으로는 어느 일정한 수준 이상으로 예측 정확도를 높이는 데(또는 오차율을 줄이는 데) 한계가 있다는 것을 의미한다. 이런 단일 알고리즘의 한계를 극복하기 위한 것이 Hybrid 모델이다. Hybrid 모델은 서로 다른 알고리즘을 통합하여, 좀 더 나은 예측 정확도를 기대하는 방법이다. 즉, 한 개의 단일 알고리즘으로는 예측율(또는 정확도)을 향상시키는데 분명한 한계를 가지기 때문에 다른 접근을 하는 새로운 알고리즘을 이용하여 예측율을 높이겠다는 것이 Hybrid 모델의 목적이다. 또한 유사한 개념 중 Combined 모델이 있다. 이것은 일반적으로 데이터 마이닝의 지도학습에서는 데이터를 크게 훈련용(Training) 데이터와 시험용(Test) 데이터로 나누는데 이 중 훈련용(Training) 데이터의 데이터 분할, 가중치 조정 등을 통해 동일한 알고리즘을 다수 적용하는 기법을 의미하며, 앙상블(Ensemble) 기법이라고도 부른다. 이런 Hybrid 모델과 Combined 모델은 특히 오차율이 기법들 간의 독립적인 경우 보다 유용한 모델링 방법으로 알려져 있다[12]. 본 논문에서는 목적 변수가 이진 변수(binary variable) 형태인 경우, 예측율을 높이는 새로운 Hybrid 모델을 제시한다. 먼저 2장에서는 기존의 Hybrid 모델과 Combined 모델을 이용한 예측율과 정확도의 성능 개선과 관련된 연구들을 정리한다. 또한 Combined 모델 중 Voting 개념, Brieman이 제시한 Bagging[6], 그리고 Freund와 Schapire가 제시한 Boosting[10]방법과 이들의 관계에 대하여 설명한다. 다음 3장에서는 본 논문에서 제시하는 오차패턴 모델 기반의 Hybrid 알고리즘을 제시한다. 본 논문에서 제시하는 Hybrid 알고리즘에서는 훈련용 데이터 집합을 2개로 분리하고 서로 다른 데이터마이닝 기법을 통해 나온 결과의 오차 부분만 다시 추출하여 2차적으로 다시 데이터마이닝 기법을 이용하여 각각 사례별로 더 잘 맞추는 기법을 판별하고, 이에 대한 오차패턴 모델을 만들어 예측율을 높이는 새로운 Hybrid 알고리즘을 제안하고자 한다. 3장에서는 본 방법의 장점 및 시사점을 정리하고, 마지막 4장에서는 여러 가지 데이터 집합을 사례로 하여, 새로운 오차패턴 모델이 들어간 Hybrid 알고리즘의 예측력 및 정확도를 기존의 분석 기법과 비교하여 살펴보도록 한다. 그리고 5장에서는 결론을 제시한다.

2.관련연구

2.1 Hybrid 기법 또는 Combined 방법을 이용한 성능 향상

지도학습 형태의 데이터 마이닝은 목적변수가 있고, 정답을 가진 훈련용 데이터를 이용하여, 패턴을 인식한 다음 그것을 검정용 또는 실제 예측할 데이터에 응용하여 그 예측력 및 정확도를 알아내는 과정을 의미한다. 이러한 지도학습 데이터 마이닝에는 신경망과 의사결정나무 분석 그리고 선형, 로지스틱 회귀분석 등이 대표적으로 사용된다. 그 외 많이 사용되는 데이터 마이닝 기법들로는 군집분석을 위한 K-평균 알고리즘과 SOM(Self-Organized Map), 그리고 연관성 분석 및 유전자 알고리즘(Genetic algorithm) 등이 있다. 성능향상을 위해서 앞에서 제시된 여러 가지 데이터 마이닝 방법을 혼합한 Hybrid 모델 및 Combined 모델도 개발되었다. 이런 Hybrid 모델 및 Combined 모델에 대한 연구로는 먼저 어떤 한 분석 기법

을 이용하여, 주요한 설명 변수를 추출하거나, 1차적으로 가지치기를 한 다음 다른 분석을 이용하여, 최종 모델을 만들어내는 경우가 있다. 예를 들어, Coenen[8] 등은 C5.0 기법을 통해서 먼저 주요 분류자(Classifier)를 발견하고, 다음 사례기반 추론(Case-based reasoning) 기법을 이용하여 분석하는 Hybrid 모델을 제시하여, Direct Mail에서 응답률 향상 분석에 적용하였다. Carvalho[7] 등은 의사결정나무 기법(C4.5)를 이용하여, 데이터를 분리한 다음, 유전자 알고리즘(Genetic algorithm)을 이용하여 최종 규칙을 찾아내는 Hybrid 방법을 제시하였다. 또한 Renpu 등은[16] 1984년 Pawlak이 제안한 Rough set 이론을 통해서, 불확실한 데이터의 정보를 어느 정도 보완한 다음 신경망을 이용하여, 최종적인 분류 규칙의 효율을 향상시키는 방법을 제시하였다. 또한 이극노[4] 등은 먼저 의사결정나무 분석 기법인 C4.5를 통해서 주요한 설명변수를 도출해 내고, 다음 이를 이용하여 신경망 분석을 하는 방식으로 2개의 서로 다른 모델을 결합하여, 이를 이동통신 고객분류에 적용하여 기존의 모델보다 더 좋은 성과를 나타냄을 보였다. 강문식[1] 등은 경쟁학습 모델과 신경망의 Back-propagation 알고리즘을 Hybrid한 HACAB[1]을 제안하였는데, 이 연구에서는 먼저 경쟁학습 모델에서 가중치를 도출해 내고, 이를 신경망의 Back-propagation 알고리즘에 적용하여, 평균 8% 정도의 패턴 분류율의 향상을 이루었다. 앞에서 제시한 Hybrid 모델들이 주로 단계별로 2개 이상의 모델을 적용하여, 결과를 향상시키려는 연구들이라면, 2개의 모델 결과의 혼합을 통해서, 새로운 예측값을 만들어 효율을 높이는 Hybrid 모델에 대한 연구도 진행이 되었다. Lin[17] 등은 일반적인 다변량 통계분석 기법과 인공지능 기법을 결합하여 도산 가능성이 높은 기업을 예측하는 연구를 수행하였으며, Conversano[9] 등은 여러 개의 통계분석 방법(회귀분석, 판별분석, 비모수통계방법, C&RT, 등)을 통해서 나온 모수들을 결합하여 다시 새로운 혼합 모델(Mixture Model)을 만들어 성능을 향상시키는 연구를 수행하였다. 또한 Versace[20]등은 금융 산업의 예측 모델에서 신경망과 유전자 알고리즘을 결합한 모델의 유용성에 대하여 연구를 수행하였다. 국내에서도 김진성[2]은 지도학습 기법인 퍼지 신경망과 비지도학습기법인 연관성 분석의 패턴을 이용한 Hybrid 메커니즘에 대한 연구를 수행하였다. 이와 같이 전혀 다른 2개의 방법을 결합하는 Hybrid 방법 이외에도, 하나의 방법을 다양하게 결합한 Combined 모델에 대한 연구도 있었다. Zhou[21] 등은 신경망 자체의 여러 결합모델이 왜 단독으로 사용된 신경망 모델보다 좋은 지를 분산의 편의(bias) 감소를 통해서 증명하였으며, Indurkhya[14] 등은 의사결정나무 기법을 여러 번 재 샘플한 다음 이를 결합하는 Voting 방법을 이용하여, 의사결정나무의 최종적인 노드들의 이익(Gain) 추정값의 향상을 연구하였다. 이러한 Hybrid 모델이나 Combined 모델이 기본적으로 단일 모델을 사용한 경우에 비해 예측율이나 분류율이 어느 정도 향상되는지에 대한 연구도 진행되어왔다. Kuncheva[15] 등이 Hybrid 모델을 이용하여, 정확도의 향상이 된 사례를 연구하였으며, Suh[19]등은 기법 간에 서로 상관성이 낮은 경우의 Combined 모델이 더욱 더 성능 향상 정도가 크다는 실증연구 결과를 RFM과 신경망, 로지스틱 회귀모델 등을 결합하여 연구한 바 있다. 이군희[3]는 전체적인 앙상블 방법에 대하여 모델의 성능 향상을 위해서 여러 모델의 의견을 종합하여, 최종 결론을 내리는 모델이 기존 모델보다 우수하다고 보편적으로 얘기할 수 있지만, 이론적인 근거는 아직 제시하지 못하고 있다고 언급하였다.

이 외에 Hybrid 모델과 Combined 모델에 대하여 다른 시각으로 접근한 연구로서 Hsu[13] 등과 같이 데이터에서 범주형 변수의 경우에는 연관성 분석을 한 다음 최적의 트리(Tree) 모델을 찾아내고, 연속형의 경우 바로 트리 모델을 이용하여 다음 단계에서 이들을 혼합한 Hybrid 모델로 학생들의 학습능력 향상에 적용한 연구를 수행하기도 하였다. 또한 Grzymala-Busse[11]와 같이 연속형 데이터를 군집분석 등을 이용하여, 이산화하여 규칙의 분류율을 높이는 또 다른 시각에서의 방법을 사용하기도 하였으며, 이재식[5] 등은 본 논문에서 제안하고자 하는 모델과 유사하게 입력자료 판별에 의한 성능 개선 방안에 대하여 연구를 하였다. 이재식 등의 논문에서는 판별모델, 기본모델, 지원모델이라는 개념을 사용하여, 특정한 분석기법으로 특정 사례에 기본모델을 적용하는 것이 좋을지 지원모델을 적용하는 것이 좋을지를 판별하고, 판별한 결과에 따라 모델을 선택하여 적용하는 방법을 사용하여, 일반 모델의 결과보다 6~7%의 정확도 향상을 가져오는 결과를 제시하였다.

2.2 예측로직과 Predictor

먼저 본 논문에서 사용할 주요 개념과 표기법을 제시하도록 한다. 본 논문에서 Predictor는 지도학습 형태의 데이터마이닝 기법에 의해서 생성된 예측 로직(logic)으로 정의한다. 즉, 데이터 마이닝의 지도학습 모델에서 먼저 훈련용 데이터 집합을 이용하여, 신경망 분석이나 의사결정나무 분석을 수행하면 일종의 예측 로직이 나오게 된다. 예를 들어 설명 변수가 성별(sex)과 소득(income)이고, 목적 변수가 이탈유무(CHURN)인 경우 신경망 분석은 훈련용 데이터를 통해서 다음과 같은 형태의 예측 로직을 생성할 수 있을 것이다.

이탈유무 예상 확률값 = 0.2 * 성별의 표준 변환값 + 0.3 * 소득수준의 표준 변환값

(계수 값은 임의로 정한 것으로 특별한 의미는 없음) 이 경우 성별의 표준 변환값이 남자이면 0.5이고, 여자이면 1인 경우, 새로운 사례가 남성이면서 소득수준의 표준 변환 값이 2인 경우 이탈 예상 확률 값은 0.2*0.5 + 2*0.3 = 0.1 + 0.6 = 0.7, 즉 70%의 이탈 가능성을 가지게 된다. 이 이탈확률을 계산하기 위해서 사용된 위 식을 이 신경망 모델의 Predictor라고 한다. 또한 의사결정나무 분석의 경우 만들어진 예측 로직에서 첫번째 규칙은 성별 "남자" 그리고 소득이 "100만원 이하"이면 "이탈", 두번째 규칙은 성별 "여자" 그리고 소득이 "500만원 이하"면 "유지" 등과 같은 여러 개의 규칙이 생성되었다고 하면, 새로운 사례가 오는 경우, 이 규칙들에 적용시켜 새로운 사례의 이탈 유무를 예측하게 된다. 이와 같이 지도학습기법에서, 예측값을 산출하기 위한 로직(logic)인 예측 로직을 본 논문에서는 Predictor라고 부르기로 하고, 보다 엄밀하게 다음과 같이 정의한다. Predictor ψ 가 데이터마이닝 기법 A와 훈련용 데이터 집합 L을 사용하여 생성되었다면, Predictor ψ 는 시험용 사례의 설명변수 벡터 집합에서 정의되고, 0과 1사이의 실수값(real number)을 반환하는 함수로 정의되고, $\psi_A(\cdot|L)$로 표시한다. 즉 i번째 사례의 설명변수 벡터 x_i에 대한 Predictor값은 $\psi_A(x_i|L)$라고 표기되고, 이는 0과 1사이의 값을 갖는다.

2.3 Hybrid 모델과 Combined 모델 그리고 데이터 마이닝

앞서도 언급했다시피 Hybrid 모델과 Combined 모델의 차이는 일반적으로 전혀 다른 분석 기법을 융합하여 사용하는 경우 Hybrid 모델이라고 하고, 동일 기법 내에서 여러 개의 데이터 집합이나 가중치를 변경하여, 여러 개의 모델을 만들어 이를 결합하는 것을 Combined 모델이라고 정의할 수 있다. 두 모델간의 의미는 조금 다르지만 이 두 방법의 공통된 의미와 목적은 Michile[23]등이 주장한 바

1) Hybrid Algorithm Combining a Competition Learning Model and BP Algorithm

와 같이 현존하는 학습 알고리즘들의 경험적 비교결과 각각의 알고리즘은 어떤 선택적인 장점(selective superiority)을 가지고 있다는데서 찾을 수 있을 것이다. 즉, 오분류가 되는 사례에 대하여, 하나의 기법 내 다양한 로직(규칙)을 발견(Combined 모형의 사상)하거나, 완전히 다른 모델의 새로운 알고리즘이 결합하여(Hybrid 모델의 사상), 오분류를 줄이고 정확도를 높이는 최종 데이터 마이닝 모델을 생성한다는 것이다.

3. 오차패턴 모델링을 이용한 Hybrid 기법

3.1 오차패턴 모델형을 이용한 Hybrid 모델의 개념

본 장에서는 제시하고자 하는 오차패턴 모델을 이용한 Hybrid 모델을 기술하도록 한다. 예를 들어, 지도학습 알고리즘을 이용하는 데이터 마이닝에서 어떤 분석 데이터가 10건이 있고, 어떤 A기법(예를 들어 의사결정나무 분석 기법)을 이용하여 모델을 만들었다고 가정하자. 그리고 그 모델을 시험용 데이터를 이용하여 검정한 결과 정확도가 70%(7건) 그리고 오분류 또는 예측을 잘못한 것이 30%(3건)였다고 가정하고, 다음 똑같은 데이터를 이용하여, 다른 모델링 기법인 B기법(예를들어 신경망 분석 기법)을 이용하여 똑같이 시험용 데이터를 이용하여 검정한 결과 역시 정확도가 70%(7건) 그리고 예측을 잘못한 것이 30%(3건)이었다고 가정한다. 이렇게 단순한 요약 정보만 있다면, 이 2개의 모델은 동일한 성능을 가졌다고 판단할 수 있다. 하지만 다음 <그림 1>과 같은 경우를 생각해 본다.

번호	실제값	A방법예측	B방법예측	Hybrid기법
1	Yes	Yes	Yes	Yes
2	Yes	Yes	Yes	Yes
3	Yes	**No**	Yes	Yes
4	No	No	No	No
5	No	No	No	No
6	Yes	Yes	Yes	Yes
7	Yes	Yes	Yes	Yes
8	Yes	**No**	**No**	**No**
9	No	No	**Yes**	No
10	No	**Yes**	**Yes**	**Yes**

<그림 1> Hybrid 기법의 예측값 선택 과정

<그림 1>에서 보면 A방법 예측과 B방법 예측은 정확도가 동일하게 70% 이지만 서로 동시에 틀린 부분도 있고(번호 8, 10번), A방법이 잘 맞춘 경우(번호 9번), B방법이 더 잘 맞춘 경우(번호 3번)도 있다. 즉, 이것을 다시 말하면 공통적으로 예측을 잘 하는 사례가 있는 반면, 공통적으로 못하는 부분도 있으며, 특정한 방법에 따라 잘 맞출 수 있는 사례와 못 맞추는 사례가 존재한다는 것이다. 이렇게 공통적으로 다 잘 맞춰줄 수 있는 부분을 제외하고 오류가 난 부분 중 각각의 방법이 잘 맞추는 경우만 가지고 오는 Hybrid 기법이 있다면, <그림 1>의 가장 우측의 결과처럼 정확도는 80%로 올라가게 될 것이다. 본 연구에서 제시하는 오차패턴 모델이란 서로 다른 2개 이상의 기법을 동일한 데이터에 적용하여, 2개 이상의 모델이 서로 다른 결과를 내는 경우만 추출하여, 데이터 집합을 구성하고, 이 데이터 집합을 가지고, 다시 A방법과 B방법이 잘 맞추는 오차패턴 모델을 생성한 다음, 실제 적용할 데이터 집합에서는 각 사례에 대하여 <그림 1>의 Hybrid 기법 결과와 같이 A방법과 B방법이 서로 잘 맞추는 사례를 맞추게 하여, 최종적으로는 오분류 및 잘못된 예측이 적은 Predictor를 만들어 내는 방법이다.

3.2 오차패턴 모델을 이용한 Hybrid 기법의 과정

구체적으로 오차패턴 모델링을 이용하여 Hybrid 모델을 만드는 과정을 정리하면 다음과 같다.

(1) 전체 훈련용 데이터 집합을 $L = \{(y_n, x_n), n = 1, 2, ...N\}$ 이라고 한다. 여기서 y_n은 목적 변수를 의미하고, x_n은 설명 변수 벡터를 의미하며, N은 데이터의 레코드 수를 의미한다. 또한 데이터 집합 L에서 목적 변수값을 가진 컬럼을 T_L로 표현한다. 그리고 훈련용 데이터를 통해서 나온 로직을 검정하기 위한 시험용(Test) 데이터를 $L_t = \{(y_t, x_t), t = 1, 2, ...T\}$ 라고 정의하도록 한다.

(2) 전체 훈련용 데이터 집합을 임의 추출방법을 이용하여, 2개로 분리를 한다. 2개의 데이터 집합을 다음과 같이 정의한다.

$$L_1 = \{(y_m, x_m), m = 1, 2, ...M\},$$
$$L_2 = \{(y_p, x_p), p = 1, 2, ...P\}$$
$$\text{단, } M + P = N$$

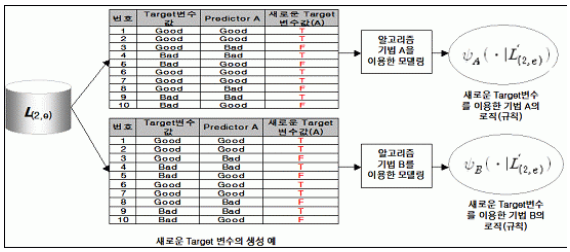
(3) 먼저 L_1 데이터 집합을 이용하여, 분석 기법 A를 이용하여, 모델링을 수행한다. 이 때 모델링을 통해서 생성한 기법 A의 로직, 즉, Predicotr를 $\psi_A(\cdot|L_1)$이라고 한다. 다음 동일한 데이터 집합에 A와는 다른 분석 기법 B를 이용하여 모델링을 수행한다. B기법을 이용하여 모델링을 수행하여 나오게 되는 로직을 $\psi_B(\cdot|L_1)$이라고 한다.

(4) 다음으로 단계 (2)에서 분리한 또 다른 훈련용 데이터 집합인 L_2에 L_1 데이터를 이용하여 생성된 두 기법의 예측 로직인 Predictor $\psi_A(\cdot|L_1)$과 $\psi_B(\cdot|L_1)$를 적용시킨다. 먼저 기법 A를 적용시켜서 나온 예측 결과(이 결과는 하나의 컬럼 형태가 될 것이다.)를 $T_{(A,L_2)}$라고 하자. 마찬가지로 기법 B를 적용시켜서 나온 결과를 $T_{(B,L_2)}$이라고 하자. 지금까지의 과정을 그림으로 표현하면 다음 <그림 2>와 같다. (5) 데이터 집합 L_2를 통해서 나온 2개의 결과 값을 서로 비교하여 결과 값이 서로 틀린 데이터 집합만을 추출한다. 이것은 <그림 1>에서 번호 3번, 9번 데이터만 추출하는 것과 동일하다고 할 수 있다. 이렇게 추출해 낸 데이터 집합을 $L_{(2,e)}$라고 정의한다. 즉, $L_{(2,e)} = \{(y_i, x_i)|(y_i, x_i) \in L_2 \text{ and} \psi_A(x_i|L_2) \neq \psi_B(x_i|L_2)\}$ 이다.

(6) 다음 데이터 집합 $L_{(2,e)}$에서 $x_i \in L_{(2,e)}$의 목적 변수 y_i값과 기법 A를 이용하여, 생성된 Predictor $\psi_A(x_i|L_1)$에 의하여 나온 결과 값과 비교하여, 서로 일치하면 T 아니면 F인 새로운 목적 변수를 생성한다. 이렇게 새롭게 파생된 목적 변수를 $T_{(A,L_{(2,e)})}$ 라고 한다. 다음 반대로 역시 기존의 목적 변수와 기법 B를 이용하여, 생성된 Predictor $\psi_B(x_i|L_1)$의 결과 값과 비교하여, 서로 일치하면 T 아니면 F인 새로운 목적 변수를 생성한다. 이렇게 새롭게 파생된 목적 변수를 $T_{(B,L_{(2,e)})}$라고 한다.

(7) $L_{(2,e)}$ 데이터 집합에서 기존의 목적 변수 $T_{L_{(2,e)}}$대신에, 새롭게 만들어진 목적 변수를 대체하여, 이 데이터 집합을 $L'_{(2,e)}$라 하고 이 데이터 집합에 기법 A, B를 다시 적용한다.

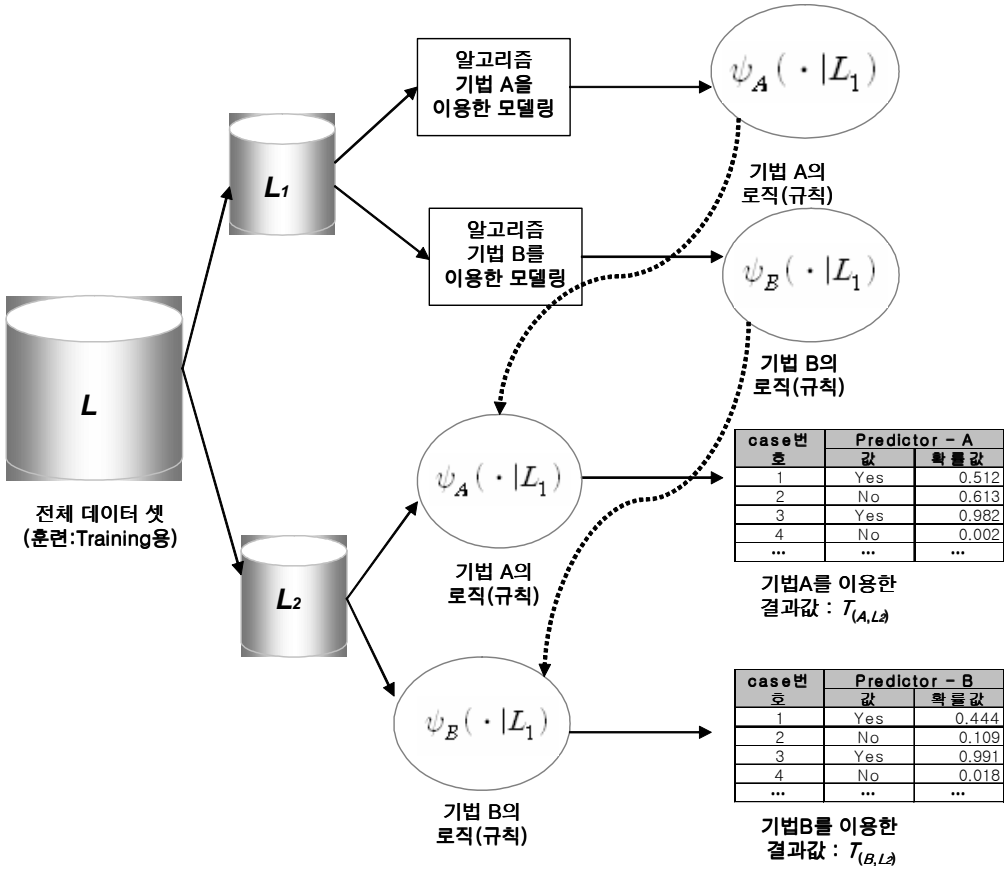
즉, $L_{(2,e)}$데이터 집합에서 먼저 목적 변수를 $T_{(A,L_{(2,e)})}$로 교체한 다음 다시 모델링 기법 A를 다시 수행하고, 다시 한 번 역시 기존의 목적 변수 대신에 $T_{(B,L_{(2,e)})}$로 교체한 다음 다시 모델링 기법 B를 수행한다. 먼저 목적 변수를 $T_{(A,L_{(2,e)})}$로 해서 모델링 기법 A를 수행한 후 발생하는 Predictor를 $\psi_A(\cdot|L'_{(2,e)})$라고 하고, 마찬가지로 $T_{(B,L_{(2,e)})}$를 목적 변수로 해서 모델링 기법 B를 수행한 후 발생하는 로직을 $\psi_B(\cdot|L'_{(2,e)})$라고 한다. 이를 예시적으로 표시한 것이 <그림 3>이다.



<그림 2> 오차 패턴 모델링의 과정

이 단계에서 만들어진 로직 $\psi_A(\cdot|L'_{(2,e)})$와 $\psi_B(\cdot|L'_{(2,e)})$의 의미는 2개의 기법 A와 B가 서로 다른 결과를 낸 데이터만 모아둔 $L_{(2,e)}$ 데이터 집합에서, 기법 A와 B가 서로 잘 맞추는 형태의 데이터 패턴을 다시 파악하는 로직이라고 할 수 있으며, 본 Hybrid 오차패턴 모델의 핵심이라고 할 수 있다. 본 논문에서는, 이 로직 $\psi_A(\cdot|L'_{(2,e)})$와 $\psi_B(\cdot|L'_{(2,e)})$을 오차패턴 모델(Error Pattern Model) 또는 오차 모델(Error Model)이라고 정의한다.

(8) 다음 이렇게 오차패턴 모델(또는 오차 모델)을 구했으면, 이를 적용한 최종 Predictor를 생성하게 되는데 이 과정은 Voting 방법을 이용한다. 예를 들어서 시험용 데이터 집합인 $L_t=\{(y_t,x_t),t=1,2,....,T\}$에서 먼저 $\psi_A(\cdot|L'_{(2,e)})$의 로직을 적용하여, 예측값이 T가 되는 사례는 $\psi_A(\cdot|L_1)$로직을 이용한 결과 값을 선택하고, 아닌 것은 $\psi_B(\cdot|L_1)$로직을 이용한 결과 값을 선택한다. 이렇게 판별하여 생성된 Predictor를 $\psi_{(A,B)}(\cdot|L\ )$ 라고 정의하고, 다음 반대로 $\psi_B(\cdot|L'_{(2,e)})$의 로직을 적용하여, 예측값이 T가 나온 사례에는 $\psi_B(\cdot|L_1)$로직을 적용한 결과 값을 선택하고 F인 것은 $\psi_A(\cdot|L_1)$로직을 적용한 결과 값을 선택한다. 이렇게 조합을 통해서 나온 최종 Predictor를 $\psi_{(B,A)}(\cdot|L\ )$라고 한다. 이들이 각기 다를 수 있으므로 안정적인 Predictor를 생성하기 위해 이 2개의 계산된 Predictor들에게서 확률값이 큰 쪽을 선택하여 만들어낸 최종 Predictor인 $\psi_{<A,B>}(\cdot|L\ )$를 생성하게 되면, 모든 과정이 완료된다. 본 내용을 예제 데이터를 사용하여 표현하면 다음 <표 1>, <표 2>와 같다.



<그림 3> 오차 패턴 모델에서 초기 데이터 분할과 두가지 기법의 적용

<표 1> 시험용 데이터에서 오차패턴 모델의 판별 결과와 두 기법의 Predictor 예제

번호	$\psi_A(\cdot \mid L'_{(2,e)})$	$\psi_B(\cdot \mid L'_{(2,e)})$	$\psi_A(\cdot \mid L_1)$		$\psi_B(\cdot \mid L_1)$	
			값	확률값	값	확률값
1	T	T	Good	0.5	Good	0.4
2	T	F	Bad	0.6	Bad	0.5
3	F	T	Good	0.7	Bad	0.5
4	F	F	Bad	0.8	Bad	0.6

<표 2> 두 기법을 Voting한 결과

번호	$\psi_{(A,B)}(\cdot \mid L)$		$\psi_{(B,A)}(\cdot \mid L)$		$\psi_{<A,B>}(\cdot \mid L)$	
	값	확률값	값	확률값	값	확률값
1	Good	0.5	Good	0.4	Good	0.5
2	Bad	0.6	Bad	0.6	Bad	0.6
3	Bad	0.5	Bad	0.5	Bad	0.5
4	Bad	0.6	Bad	0.8	Bad	0.8

3.3 오차패턴 모델을 이용한 Hybrid 기법의 의미

오차패턴 모델을 이용한 Hybrid 기법을 만들어 내는 가장 큰 목적은 더 높은 예측 정확도를 위해서 다른 기법들 간에 자신의 기법이 더욱 효율적인 데이터 사례에만 해당 기법이 적용되도록 하는 것이다. 이를 위해 오분류 데이터에 대해 데이터 마이닝 모델링을 한 번 더 사용하여, 해당 기법이 잘 맞추는 데이터는 어떤 데이터인가를 구분하는 패턴을 알아내는 모델링 과정이 추가가 된 것이다. 이 경우 일반적으로 2개의 기법보다는 조금이라도 데이터의 오분류 가능성이 낮아지고, 적어도 두 기법 중 더 나은 기법 정도의 정확도를 제공하는 구조여서, 상당히 안정성이 있다고 논리적으로는 설명할 수 있다. 그러나 그에 비해 모델링하는 단계가 복잡하고, 한 번 더 모델링을 수행함에 따라 수행시간이 더 경과되는 단점도 있을 것이다. 다음 장에서는 실제 여러 데이터 집합에 본 방법을 적용한 실험 결과에 대하여 설명한다.

4. 성능 비교를 위한 실험 결과

4.1 실험 가정

제시한 오차패턴 모델의 성능 비교를 위한 실험에서는 다음과 같은 몇 가지 가정을 두었다.

첫째, 실험 데이터의 목적 변수는 전부 이분형(Binary) 형태이다. 예를 들면 Yes/No, Response/No Response, T/F, Good/Bad 등의 형태이다.

둘째, 사용된 기법은 신경망의 MLP 방법과 의사결정나무 분석 기법 중 Quinlan이 개발한 C5.0 기법 2가지만을 활용하였다.

셋째, 사용되는 2개의 기법은 데이터에 따라 정확도 향상을 위한 각종 option을 변경하지 않는다. 이는 option 변경으로 인해 정확도가 향상되는 부분을 배제하기 위해서이다.

MLP와 C5.0을 이용한 것은 신경망과 의사결정나무 분석은 매우 이질적인 기법이므로 이를 통해서 Hybrid 모델의 사상을 잘 나타낼 것이기 때문이다. 본 실험의 수행을 위해서, 알고리즘을 자체적으로 프로그래밍하는 경우 그 신뢰성을 보장할 수 없기때문에 전 세계적으로 광범위한 활용과 신뢰도를 가진 SPSS Inc.[2]사의 데이터 마이닝 S/W인 클레멘타인 8.1(Clementine 8.1)을 이용하여 분석을 수행하였다.

4.2 실험 데이터의 설명

본 실험에서는 서로 다른 10가지 종류의 데이터 집합을 활용하였으며, 데이터 집합의 간단한 설명은 다음의 <표 3>과 같다.

<표 3> 실험에 사용된 데이터 집합의 설명

번호	설명변수	목적변수	비고
1	범주형 : 5 연속형 : 13	가입 해지 여부	한국의 유선전용 통신사 자료
2	범주형 : 8 연속형 : 6	소득 5만달러 이상/이하	소득예상계층 분류자료
3	범주형 : 7 연속형 : 3	반응/비반응	미국의 잡지구독 캠페인 자료
4	범주형 : 6 연속형 : 3	이탈/유지	이동통신사의 가입자 해지 자료
5	범주형 : 6 연속형 : 7	이탈/유지	영국의 유선 통신사 해지 자료
6	범주형 : 3 연속형 : 10	가입/비가입	보험사의 종신보험가입 자료
7	범주형 : 3 연속형 : 11	우수/비우수	증권사의 우수고객 여부 자료
8	범주형 : 9 연속형 : 14	구매/비구매	홈쇼핑의 특정제품 구매여부
9	범주형 : 10 연속형 : 10	구매/비구매	영국의 할인점 중 특정제품 구매여부
10	범주형 : 6 연속형 : 31	이탈/유지	미국의 유선통신사의 유선전화 가입여부

4.3 실험 결과

위의 <표 3>에서 나열한 10개 서로 다른 데이터 집합에 대하여 신경망, C5.0 그리고 오차패턴 모델을 이용한 Hybrid 모델을 수행하였을 때의 정확도와 그에 관련된 데이터들을 정리한 것이 <표 4>와 <표 5>이다.

<표 4> 실험 데이터의 기본 현황

데이터 번호	전체 데이터수	시험용 데이터수	C5.0/신경망 불일치 비율
1	14,490	4,765	1,500 (31.48%)
2	32,561	10,750	1,017 (9.46%)
3	12,000	3,995	227 (5.68%)
4	3,284	1,079	258 (23.9%)
5	4,574	1,539	105 (6.82%)
6	100,000	33,013	1,314 (6.82%)
7	228,774	76,243	15,194 (19.93%)
8	2,181	754	236 (31.29%)
9	30,000	10,000	4,853 (48.53%)
10	16,004	5,396	296 (5.48%)

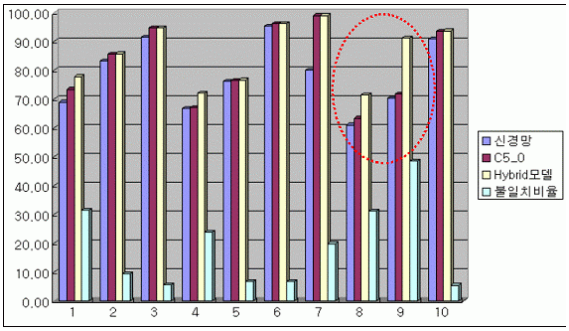
2) SPSS Inc. (http://www.spss.com), SPSS Korea (http://www.spss.co.kr)

<표 5> 실험결과

데이터 번호	신경망 단독정확도	C5.0 단독 정확도	**오차패턴 모델을 이용한 Hybrid 모델 정확도**
1	68.96% True : 3,286 False : 1,479	73.33% True : 3,494 False : 1,271	**77.73% (향상도: 3.40%) True : 9,209 False : 1,541**
2	83.23% True : 8,947 False : 1,803	85.45% True : 9,186 False : 1,564	**85.67% (향상도: 0.22%) True : 9,209 False : 1,541**
3	91.34% True : 3,649 False : 346	94.67% True : 3,782 False : 213	**94.67% (향상도: 0%) True : 3,782 False : 213**
4	66.73% True : 720 False : 359	67.10% True : 724 False : 355	**72.01% (향상도: 4.91%) True : 777 False : 302**
5	76.09% True : 1,171 False : 368	76.54% True : 1,178 False : 361	**76.60% (향상도: 0.06%) True : 1,179 False : 360**
6	95.14% True : 31,409 False : 1,604	96.01% True : 31,697 False : 1,316	**96.22% (향상도: 0.21%) True : 31,765 False : 1,248**
7	80.02% True : 61,003 False : 15,231	98.87% True : 75,371 False : 863	**98.90% (향상도: 0.03%) True : 75,398 False : 836**
8	61.01% True : 460 False : 294	63.40% True : 478 False : 276	**71.49% (향상도: 8.09%) True : 539 False : 215**
9	70.30% True : 7,030 False : 2,970	71.65% True : 7,165 False : 2,835	**91.10% (향상도: 19.45%) True : 9,110 False : 890**
10	90.77% True : 4,898 False : 498	93.48% True : 5,044 False : 352	**93.64% (향상도: 1.16%) True : 5,053 False : 343**

<표 4>와 <표 5>의 결과를 보면 전체적으로 오차패턴 모델을 이용한 Hybrid 모델의 정확도가 신경망 단독 정확도 또는 C5.0 단독 정확도보다는 더 높다는 것을 알 수 있다. Hybird 모델과 신경망 C5.0간의 정확도의 차이를 반복분산분석(Repeated ANOVA)을 사용하여 비교한 결과, 10가지 경우 모두 차이가 유의 수준 0.01%에서 통계적으로 유의하였다. 그러나 정확도의 편차는 상당히 많이 나는 것도 알 수가 있다. 예를 들어 9번과 8번 데이터들의 경우에는 기존의 신경망이나 C5.0 보다 10~20% 이상의 정확도가 향상된 반면, 2번, 3번 데이터들의 경우 0~0.3% 정도 수준의 정확도 향상이 이루어 진 것을 볼 수 있다. 다음의 <그림 6>은 각 기법의 예측력과, 기법과 오분류율 간의 편차를 하나의 그래프로 표현을 한 것이다.

<그림 4>의 그래프에서 보면 전체적으로 Hybrid 모델의 예측 정확도가 다른 두 기법보다 높거나 동일한 수준을 나타내는 것을 알 수 있다. 특히 8번과 9번 데이터의 경우 상당히 높은 정확도 향상을 보여주고 있다. 이 때의 불일치 비율 역시 31.29%, 48.53%로 다른 데이터 집합의 경우보다 불일치 비율이 높은 것을 <그림 4>을 통해서도 알 수 있다. 이렇게 데이터 집합마다 정확도 향상의 편차가 많은 것은 데이터의 성격에 따라 여러 가지 요인이 있겠지만 <표 3>과 <표 4> 그리고 <표 5>를 통해서 보면, 향상도가 큰 데이터는 다음과 같은 몇 가지 특징 및 요인이 있



<그림 4> 신경망, C5.0, Hybrid 모델 및 불일치 비율의 비교 그래프

다는 것을 발견할 수 있다.

첫째 향상력이 높아지는 요인으로는 데이터 집합에서 2개의 기법 간 사례의 예측 불일치 비율이 높아야 한다는 것이다. 이것은 당연히 서로 엇갈려서 맞추는 사례가 많아야지만, 서로 간에 부족한 부분에 대한 보충을 하며, 향상도를 높여주게 된다고 이해할 수 있다. <표 5>에서 비교적 향상도가 높은 9번, 8번, 4번, 1번 데이터들의 경우 불일치 비율이 높은 경우는 약 48%부터, 낮은 경우는 약 24 %까지 두 기법간의 불일치되는 비율이 높게 나타났으나, 향상도가 낮은 5번, 6번, 10번 데이터의 경우에는 10% 이하의 불일치 비율이 나타났다. 더 생각해 볼 수 있는 것은 불일치 비율이 낮은 경우의 데이터 오류 사례들은 분석 기법만 가지고는 더 이상 향상이 어려운 데이터들이라고 설명할 수 있을 것이다.

둘째 요인으로는 데이터가 불일치 비율이 높다는 조건이 만족되면, 다음에는 데이터에서 설명 변수들의 데이터 유형이 범주형과 연속형이 적절히 균형있게 존재하고, 설명변수의 수가 많을수록 정확도의 향상도가 높아진다는 사실이다. 향상도가 높은 8번, 9번, 2번 데이터들의 경우 거의 범주형 데이터와 연속형 데이터의 설명 변수 수가 비슷비슷하게 존재를 하고, 설명 변수의 수도 비교적 많은 것을 알 수 있다. 이는 신경망 분석의 경우 알고리즘의 특성상 연속형에 좀 더 적중률이 높고, 의사결정 나무 분석의 경우 분리를 시켜야 하는 알고리즘의 특성상 범주형의 데이터에 좀 더 정확도를 높기 때문에, 이 2가지 기법을 결합한 Hybrid 모델의 정확도가 더욱 향상되는 효과를 발휘하는 것으로 생각해 볼 수 있다.

그 외의 경우로는 7번 데이터 집합의 경우처럼 어느 한쪽 기법의 정확도가 지나치게 뛰어난 경우(신경망 80.02%, C5.0 98.87%) 당연히 Hybrid 모델이 별다른 효과를 발휘하지 못하고, 또한 전체적으로 정확도가 90% 이상인 경우(6번, 3번 데이터)도 그다지 Hybrid 모델의 효과가 좋지 않음을 알 수 있다.

5. 결론

본 연구에서는 데이터 마이닝의 지도학습 모델에서 Hybrid 모델 및 Combined 모델을 이용하는데 있어 기존의 Voting이나 Bagging, Boosting 이외에, 오차 패턴 모델(Error Pattern Modeling)을 이용한 새로운 Hybrid 모델을 제시하고, 제시한 Hybrid 모델이 기존의 단일 분석 알고리즘을 사용하여 분석(모델링)했을 때보다 얼마만큼 더 많은 정확도의 향상을 가지고 오는지에 대하여 10개의 데이터 집합을 사용하여 분석을 하였다. 분석 결과를 요약하면, 본 모델의 정확도는 최소한 같거나 향상을 가져오며, 특히 기법 간의 사례의 정확도 불일치 비율이 높고, 데이터에서 설명 변수의 유형이 범주형과 연속형이 골고루 섞여 있으며, 설명 변수의 수가 많은 경우 향상도가 높은 것으로 나타났다. 따라서 분석하고자 하는 데이터를 탐색해

서 위와 같은 조건을 만족하는 데이터라면 본 모델을 사용하여 좀 더 높은 정확도를 가진 새로운 모델을 개발할 수 있을 것으로 사료된다. 또한 분석을 하는 비즈니스의 상황에서, 아주 조금이라도 정확도가 더 높은 모델을 필요로 하는 경우(예를 들면 카드사 등에서 사기적발과 같이 1건의 피해라도 매우 큰 영향을 주는 경우)에도 본 모델을 적용한다면, 좋은 성과를 올릴 수 있을 것으로 판단된다. 그러나 기법 간의 불일치 비율이 아주 낮거나, 특정한 한 기법의 정확도가 유난히 높은 경우, 그리고 전체적으로 정확도가 높은 경우에서는 본 모델의 효과가 기존 모델에 비해서 크게 향상되지 않는 것으로도 나타났다. 이런 경우에서는 모델의 수행 시간과 자원의 절약 차원에서 본 모델을 사용하지 않는 것이 더 나을 수도 있을 것이다.

추후 연구방향은 다음과 같다. 본 연구에서는 신경망과 C5.0만을 한정하여 실험하였으나, 타 지도학습 모델들을 포함하여 검토하는 것이 필요하다. 또한 목적 변수를 이분형(Binary) 데이터로 한정하였는데, 범주가 3개 이상인 범주형 데이터와 연속형으로 확장하는 노력이 필요하다. 아울러서 어떠한 특성을 갖는 데이터 집합에서 더욱 향상도가 높아지는 지에 대한 추가적인 연구도 필요하다.

<참고문헌>

[1] 강문식, 이상용, “데이터 마이닝을 위한 경쟁학습모델과 BP알고리즘을 결합한 하이브리드 신경망”, 「정보기술과 데이터베이스 저널」, 제9권 2호 (2002), pp.1-16.

[2] 김진성, “연관규칙과 퍼지 인공신경망에 기반한 하이브리드 데이터 마이닝 메커니즘에 대한 연구”, 「한국경영과학회/대한산업공학회 2003 춘계 공동학술대회 논문집」, (2003), pp.884-888.

[3] 이군희, “모형평가와 앙상블을 이용한 데이터 마이닝에 관한 연구”, 「서강경영논총」, 제9권 (1998), pp.293-306.

[4] 이극노, 이홍철, “이동통신고객 분류를 위한 의사결정나무(C4.5)와 신경망 결합 알고리즘 연구”, 「한국지능정보시스템학회지」, 제9권 1호 (2003), pp.139-155.

[5] 이재식, 이진천, “입력자료 판별에 의한 데이터 마이닝 성능개선” 「한국지능정보학회학술대회」, (2000), pp.293-303.

[6] Brieman, L. "Bagging Predictors", *Machine Learning*, Vol.24, No.2(1996), pp.123-140.

[7] Carvalho, Deborah R. and Alex A. Freitas "Hybrid Decision Tree/Genetic Algorithm Method for Data Mining" *Information Sciences,* Vol.163, No.1/3(2004), pp.13-35.

[8] Coenen, F. G. Swinnen, K.Vanhoof and G.Wets "The Improvement of Response Modeling: Combining Rule-induction and Case-based Reasoning", *Expert Systems with Application*, Vol.18, No.4(2000), pp.307-313.

[9] Conversano, Claudio, Roberta Siciliano and Francesco Mola, "Generalized Additive Multi-mixture Model for Data Mining", *Computational Statistics & Data Analysis,* Vol.38, No.4(2002), pp.487-500.

[10] Freund, Y. and Rober E. Schapire, "Experiments with a New Boosting Algorithm", *Proceedings of 13th International Conference on Machine Learning,* Morgan Kaufmann(1996), pp.148-156.

[11] Grzymala-Busse, Jergy W, "A Comparison of Three Strategies to Rule Induction from Data with Numerical Attributes", *Electronic Notes in Theoretical Computer Science*, Vol.82, No.4(2003), pp.1-9.

[12] Hansen, L.K. and P.Salaman, "Neural Networks Ensembles", *Transactions on Pattern Analysis and Machine Intelligence*, Vol.12, No.10(1990), pp.993-1001.

[13] Hsu, P. L., R.Lai, CC.Chui,and C.I.Hsu, "The Hybrid of Association Rule Algorithms and Genetic Algorithm for Tree Induction: An Example of Predicting the Student Course Performance", *Expert Systems with Application*, Vol.25, No.1(2003), pp.51-62.

[14] Indurkhya, Nitin and Sholom M. Weiss "Estimating Performance Gains for Voted Decision Trees" *Intelligent Data Analysis,* Vol.2, No.1/4(1998), pp.303-310.

[15] Kuncheva, L.I.C. Bezdek and M.A.Shutton, "On Combining Multiple Classifiers by Fuzzy Templates", *International Conference on Artificial Neural Networks IEEE*, (1998) pp193-197.

[16] Li, Renpu and Zheng-ou Wang "Mining Classification Rules Using Rough Sets and Neural Networks", *European Journal of Operational Research*, Vol.157, No.2(2004), pp.439-448.

[17] Lin, Feng Yu and Salley McClean "A Data Mining Approach to the Prediction of Corporate Failure", *Knowledge-Based Systems,* Vol.14, No.3/4(2001),pp.189-195.

[18] Michie D., D.J.Spiegelhalter, and C.Taylor. *Machine Learning, Neural and Statistical Classification,* Ellis Horwood, 1994.

[19] Suh, E.H, K.C.Noh and C.K.Suh "Customer List Segmentation Using the Combined Response Model", *Expert Systems with Application*, Vol.17, No.2(1999), pp.89-97.

[20] Versace, Massimiliano, Rushi Bhatt, Oliver Hinds and Mark Shiffer "Predicting the Exchange Traded Fund DIA with a Combination of Genetic Algorithm and Neural Networks", *Expert Systems with Application*, Vol.27, No.3(2004), pp.417-425.

[21] Zhou, Zhi-Hua, Jianxin Wu and Wei Tang, "Ensembling Neural Networks: Many Could Be Better Than All", *Artificial Intelligence,* Vol.137, No.1/2(2002), pp.239-263.